Chapter 9

Time Lags, Dummy Variables, and Transformations

This chapter describes three broad classes of refinements on multiple linear regression:

1. Lags in the timing of causal relationships among time-series variables.
2. Dummy variables to reflect known but usually unmeasured causal influences that may affect one or more individual values in time-series data, as well as levels and slopes.
3. Transformations of variables to allow flexibility in the functional form of causal relationships.

These three classes of refinement substantially increase our ability to find and express relationships among data in complex models. In the forecaster's mind, each class of refinement provides a route for handling a class of difficulties in building statistical forecasting systems in situations where linear regression is inadequate.

9.1 Economic Basis for Time Lags

The fundamental idea of a time lag is that the occurrence of one independent event will cause a second dependent event to occur after a time period of some length. This lag is essentially an economic reaction time, reflecting the response time of human decision makers in both businesses and households and the delay times through systems of making decisions to buy, order, produce, deliver, and pay for goods and services. Time lags are zero or very short for low-value nondurable goods such as food. But time lags may be as long as two or three quarters for intermediate-type durables such as automobiles and construction tractors, and may be as long as five to ten years for very large, expensive, highly durable goods, such as electric generating facilities for public utilities.

9.2 Graphic and Computer Determination of Single Lags

A simple regression equation designed to explain variations over time in the mean value of the dependent variable is given by:

$$Y_{ct} = b_1 + b_2X_{2t}; \text{ where } t \text{ is the time index} \qquad (9.1)$$

This predicting equation assumes that the current value of Y depends on the current value of X but *not* on any of the past values of X. This assumption is very limiting and can be relaxed by using single or multiple lags described subsequently.

Determination of a single time lag can be illustrated with the data on industrial wheel tractors and total construction contracts. The awarding of a given level of construction contracts can be readily visualized as the economic event that causes the purchase by construction contractors of industrial wheel tractors concurrently or one or more quarters later.

Ideally, the length of the lag should be determined from studies of the decision-making patterns of purchasers for the product in question. In the case of construction contractors, this would require interviewing an adequate random sample of U.S. contractors to determine the times between the awarding of construction contracts and their purchasing of industrial wheel tractors for use in this construction. The critical part of this information gathering would be to determine when an important change in the trend of construction, such as a downturn after a few quarters of rising contracts or an upturn after a few quarters of declining contracts, would result in a contractor's changing his rate of purchasing tractors. This study would be a complex one and has never been done.

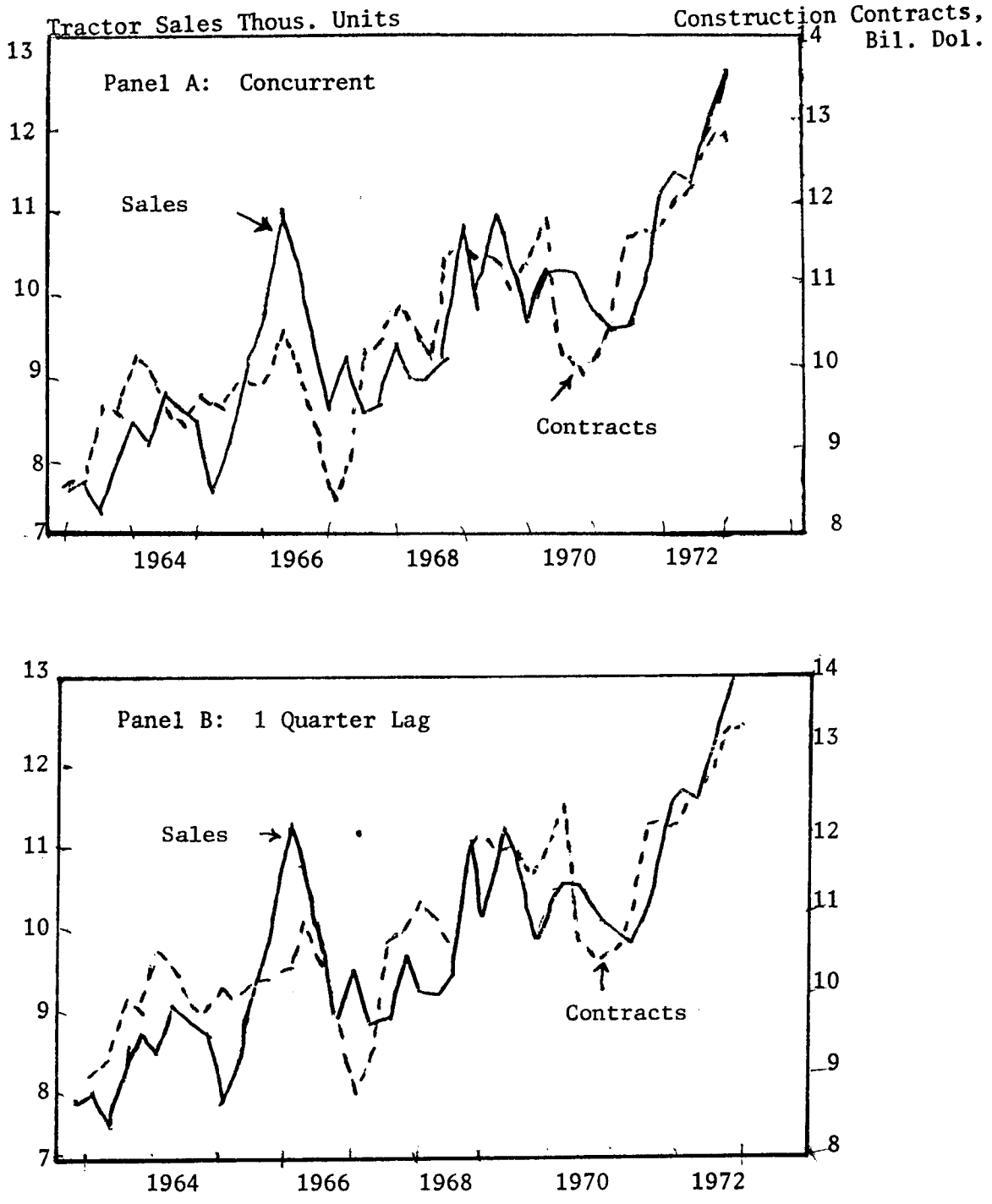
In sales forecasting work, lag relationships usually are determined by the statistical correspondence of sales and explanatory variables at the turning points of business cycles.

Graphic determination of a single lag can be done by visually matching the business cycle peaks and troughs of the two variables. You may do this by placing a chart of the sales variable over a light box or in front of a window, placing a chart of the explanatory variable on top, and then moving the explanatory variable chart to the right or left until the best overall match of timing is found. This process is illustrated in Figure 9.1, Panels A, B, C, and D, where vertical scales have been adjusted to show approximately the same level and amplitude.

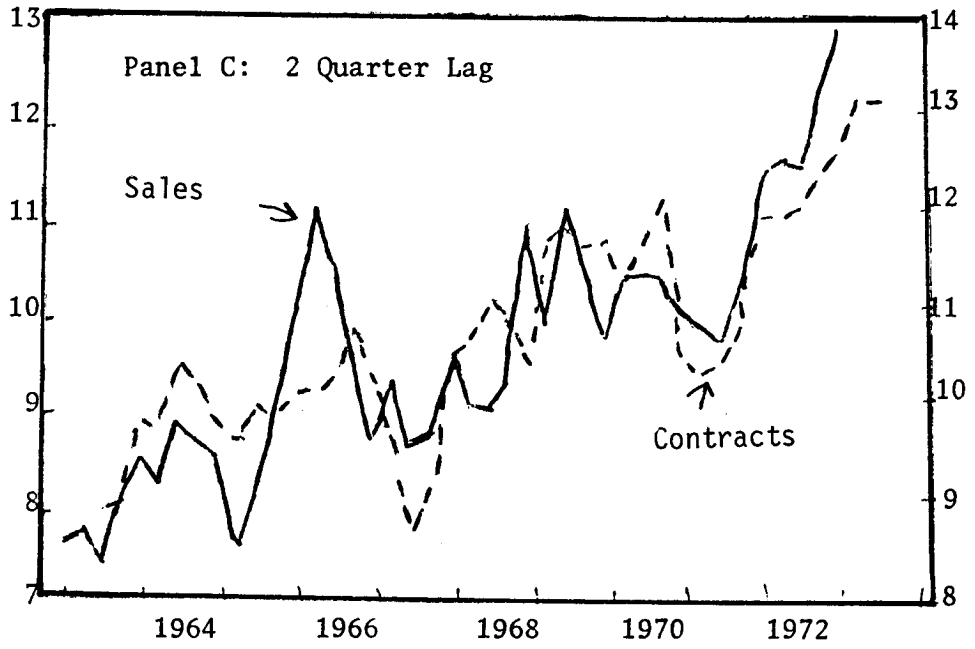
Figure 9.1

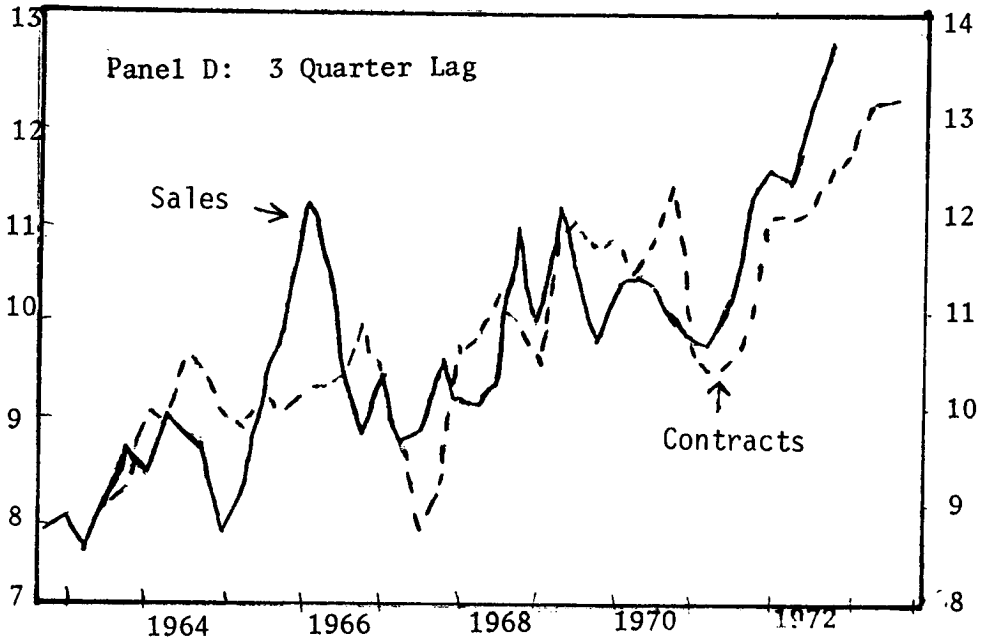
Industrial Wheel Tractor Sales and Construction Contracts,

With Sales Lagged 0 to 4 Quarters



Continued on next page


Panel C: 2 Quarter Lag
Sales
Contracts
13
12
11
10
9
8
7
14
13
12
11
10
9
8
1964
1966
1968
1970
1972


Panel D: 3 Quarter Lag
Sales
Contracts
13
12
11
10
9
8
7
14
13
12
11
10
9
8
1964
1966
1968
1970
1972

The one-quarter lag (of industrial wheel tractor sales after construction contracts) in Panel B and the two-quarter lag in Panel C appear visually to be the best, with Panel B being slightly preferred because of the better fit at the right side, for the most recent data. Relationships in Panel A and Panel D are less close.

The statistical determination of the best single lag consists of calculating the coefficients of simple linear correlation with lags of zero, one, two, three, and four quarters, and then choosing that lag with the highest coefficient of correlation. For industrial wheel tractors and construction contracts, the correlation coefficients, *r*, appear in Table 9.1.

Table 9.1
Lag of Tractor Sales After Construction Contracts, As Measured by Coefficients of Correlation, r.

Lag in Quarters	r
0	.80
1	.84
2	.83
3	.77
4	.68

Thus the one-quarter lag is best (has the highest r), but the two-quarter lag is nearly as good. Notice that a concurrent relationship (i.e., zero lag) is not a great deal worse than a one- or two-quarter lag. Most forecasting practitioners prefer to use any increase in R^2 that is through a lag relationship when the lag is justified by economic and decision-making theory, and therefore they would use the one- or two-quarter lag for tractors rather than no lag.

For a lag of one quarter, the equation form is:

$$Y_{ct} = b_1 + b_2 X_{t-1} \tag{9.2}$$

This form shows the time subscript of X as $t-1$, meaning that the value of X_t for one quarter earlier in time will be regressed on Y_t. *The dependent sales variable is always expressed in natural time, and the explanatory variables are expressed in relation to the dependent variable.* After some practice, you will automatically think of Equation 9.2 as describing a one-quarter lag of Y after X, even though the "−1" notation is on the X value.

A second example of timing differences in variables appears in Figure 9.2 with Process Control Company Sales and U.S. New Plant and Equipment Expenditures for Manufacturing, all in seasonally adjusted current dollars. Note that the peaks in Process Control Sales occur sooner in 1966 and in 1969 than the peaks in NPEE. Also, in 1971 the upturn in Process Control started sooner than the upturn in NPEE. A time lag or lead is called for in the same way that a lag was needed in Figure 9.1 for tractors and contracts.

Figure 9.3 shows the same data with NPEE plotted two quarters earlier and with much improved correspondence of the trends and peaks. The scales on these charts have been plotted to force the means of the two series to correspond, but no attempt was made to force correspondence of the cyclical amplitude of the two series.

The timing difference between these two series must be explained from an economic standpoint. Here the business cycle turns of Process Control Company sales occur about two quarters sooner than NPEE; i.e., the sales variable is correlated with a two-quarter lead of the explanatory variable. This case of sales occurring sooner than the explanatory variable is the reverse of the usual case, but it can be readily explained by the following series of actions:

1. Manufacturer decides to order New Plant and Equipment (Manufacturing), presumably based on a comparison between the desired stock of NPEE (Mfr.) and the existing stock.
2. Producer of New Plant and Equipment (Mfr.) receives the order and places orders in the plant and on outside suppliers for components.
3. Process Control sells and ships control devices as an outside supplier of components.
4. Producer of NPEE completes the order, delivers it to the Manufacturer in No. 1, and payment is recorded as an expenditure in the NPEE series.

Thus the two-quarter lead of sales relative to NPEE is simply a difference in the timing of sales of a new component versus payment for the completed machine in NPEE. The timing relationship is logical and acceptable.

Such a time lead relationship can be considered a negative lag and is shown notationally by reversing the sign of the subscript for the lag. Thus a lag of one quarter is shown as $X_{2(t-1)}$, but a *lead* of one quarter is shown as $X_{2(t+1)}$. Think of this notation as saying that data for X_2 one quarter ahead in time will be regressed with data for sales in the current (or natural) quarter. For the Process Control case, the two-quarter lead results in the following regression equation:

$$Y_{ct} = 6.21 + 0.621\ X_{2(t+2)} \tag{9.3}$$

9.3 Multiple Lags and Their Evaluation

If in the case of industrial wheel tractors and construction contracts, a one-quarter lag relationship yielded the highest coefficient of correlation, and a two-quarter lag was nearly as high, why not use a combination of two or more lagged explanatory variables? This combination is possible and usually leads to a higher R^2. A problem, however, is that the two or more explanatory variables, which really are the same variable but with different time lags, are likely to be correlated. This condition, you may recall, is multicollinearity. If multicollinearity exists among the independent variables, then the typical tests of significance of the regression coefficients, b_i, are not precise because of bias in their standard errors. But let us first calculate the multiple lag and then evaluate this problem.

Where the lags are relatively short, typically a few quarters, and where only a few lag terms are needed, we can

Figure 9.2

PROCESS CONTROL NEW ORDERS AND
NEW PLANT AND EQUIPMENT EXPENDITURES, MFR.
SEASONALLY ADJUSTED QUARTERLY DATA

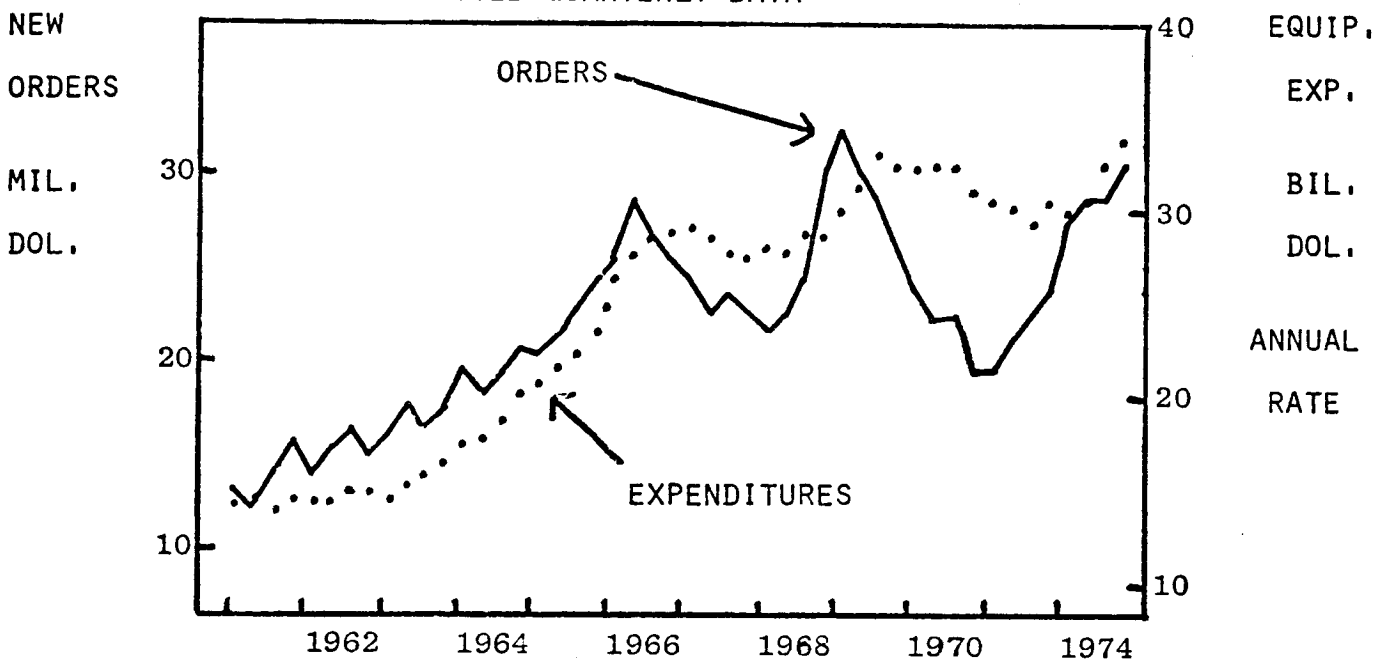
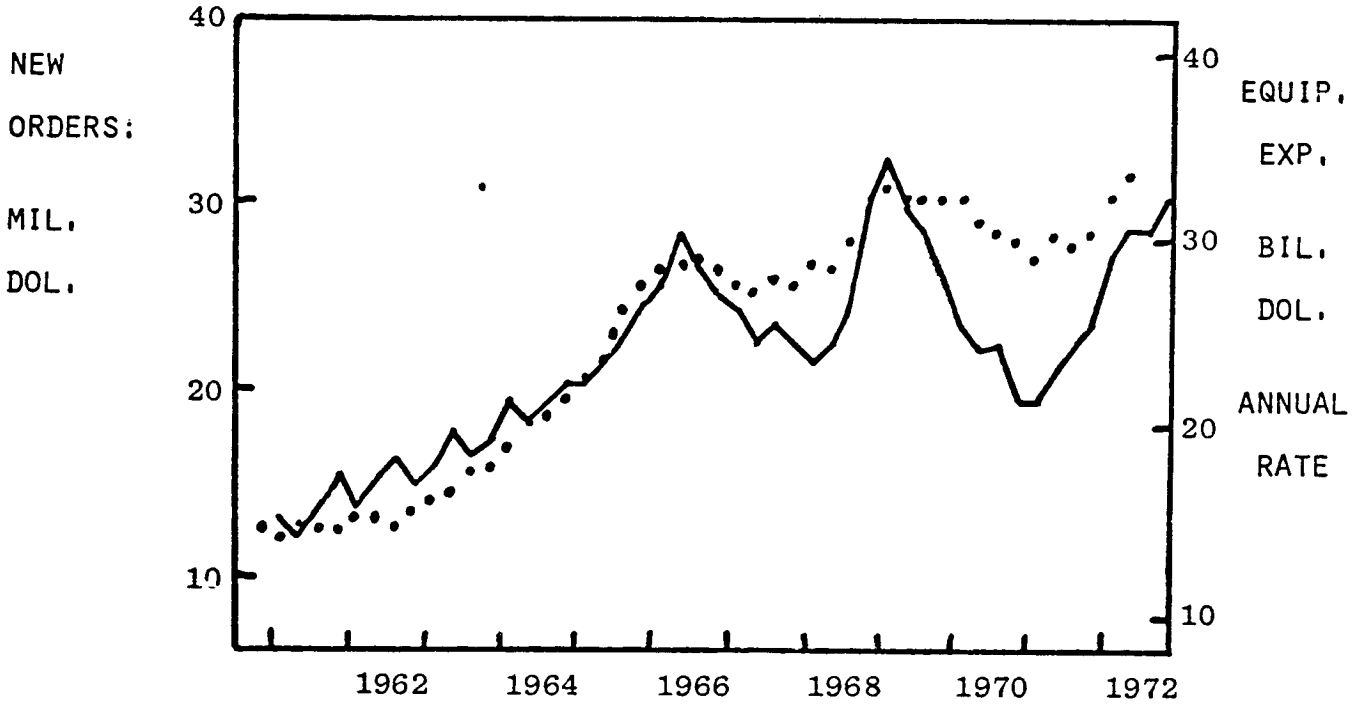


Figure 9.3

PROCESS CONTROL WITH TWO-QUARTER LEAD, AND
NEW PLANT & EQUIPMENT EXPENDITURES



use the simple approach, *multiple lags*. Multiple lags refer to one explanatory variable that has been entered with several consecutive lags, 0 to 4 quarters for example, into a multiple regression equation.

We regress industrial wheel tractors with construction contracts, X_{2t}, with five lags as X_{2t}, $X_{2(t-1)}$, $X_{2(t-2)}$, $X_{2(t-3)}$, and $X_{2(t-4)}$. Here $X_{2(t-3)}$ is the best single independent variable, as indicated by the simple correlation coefficients in Section 9.2. The regression equation with one independent variable is:

$$Y_{ct} = b_1 + b_2 X_{2(t-1)} \tag{9.4}$$

or:

$$Y_{ct} = 1.79 + 1.013\, X_{2(t-1)} \tag{9.5}$$

Standard error = .107
T ratio = 9.51
$R^2 = .70$

The t ratio is highly significant relative to the critical t of 2.42 of the 1 percent level.

Further insight can be gained by inspecting the correlation coefficient matrix in Table 9.2 for all possible combinations of two variables. The correlation coefficients in the bottom row are the same as those given in the table in Section 9.2. Let us consider adding a second lagged independent variable, such as $X_{2(t-2)}$, the variable with the second highest r. The coefficient of correlation of $X_{2(t-2)}$ with $X_{2(t-1)}$ is 0.86. Thus $X_{2(t-2)}$ does not add much "new information" after $X_{2(t-1)}$ has been used. Similarly the coefficient of correlation of X_{2t} with $X_{2(t-1)}$ is .87, so X_{2t} does not add much further information either. Intuitively then, we may think of $X_{2(t-1)}$ as representing not only itself but also the adjacent lags, X_{2t} and $X_{2(t-2)}$. The partial F test in stepwise regression yields rigorous measurement of the foregoing intuitive approach.

The best combination of two independent variables is $X_{2(t-1)}$ and $X_{2(t-3)}$. (Notice that X_{2t}, $X_{2(t-2)}$, and $X_{2(t-4)}$ are excluded.) Intuitively, we might also consider $X_{2(t-3)}$ as representing both itself and the longer lag of $X_{2(t-4)}$ since their intercorrelation is .83.

Our multiple lag equation is:

$Y_{ct} = -3.48 +$	$.709\, X_{2(t-1)} +$	$.461\, X_{2(t-3)}$	(9.6)
Std. error	.136	.147	
T ratio	5.20	3.14	

$R^2 = .761$

The correlation between the two explanatory variables on the right is +.71, which is less than our .85 rule of thumb in Chapter 8. The .71 is large enough to moderately bias the standard errors but not drastically. Thus we recognize that the classical assumption of no multicollinearity between explanatory variables in linear regression has been violated somewhat, and, therefore, the tests of significance of the regression coefficients are not entirely accurate. Given multicollinearity of only .71, however, and the substantial size of the t ratios, we judge that the two regression coefficients in Equation 9.6 remain significant.[1]

Ideally, we might like to search for additional independent variables, but extensive search has shown that "construction contracts" is the best conveniently available variable. Given the resources available in this instance, one compromise is to accept Equation 9.6 as a useful, though not ideal, forecasting equation, and to keep in mind two points: (1) the two independent variables must be forecast jointly, under the same assumptions about the economy; and (2) the t statistics cannot be individually considered as highly accurate in reflecting the significance of each variable. Notice that in this case the acceptance of $X_{2(t-1)}$ alone as an explanatory variable can be readily substantiated by a high t ratio, 9.51. Accepting the same

Table 9.2 Coefficients of Correlation: r_{ij}

Variable	Variable					
	X_{2t}	$X_{2(t-1)}$	$X_{2(t-2)}$	$X_{2(t-3)}$	$X_{2(t-4)}$	Y_t
X_{2t}	1.00					
$X_{2(t-1)}$	.87	1.00				
$X_{2(t-2)}$	.73	.86	1.00			
$X_{2(t-3)}$	.63	.71	.85	1.00		
$X_{2(-4)}$	.58	.62	.70	.83	1.00	
Y_t	.80	.84	.83	.77	.68	1.000

variable again with a different lag does not require as rigorous a test for significance as if we were testing a different economic concept for a second explanatory variable.

We analyzed Equation 9.6 for significant regression coefficients and stopped with two explanatory variables because further variables yielded much smaller regression coefficients. A further important reason for stopping with two variables is that negative regression coefficients will frequently appear if many lagged variables are included. Negative coefficients for lagged variables usually are not justified by economic causation. Such negative coefficients usually reflect multicollinearity or a calculating error and constitute a strong diagnostic signal of nonsignificant regression.

The reader may ask, "Why not use a fractional lag like 1½ quarters?" The answer is in two parts:

1. Use a fractional lag if data for shorter periods than quarters are available and if the fractional lag is better than an integral lag. In the case of industrial wheel tractors and construction contracts, both original time series are available by months. Then determine the best single lag in months and perform all regression with monthly data. A 4-month lag equals a 1 1/3-quarter lag. Monthly data would allow any single lag in multiples of 1/3 quarter. Analogously, weekly data would allow any single lag in multiples of 1/13 quarter. When using explanatory variables from the national income accounts, however, only quarterly data are available. A possibility here is to group monthly sales data into whatever quarterly arrangement yields a good single integral lag with the available quarterly explanatory variable. Data grouping then yields a fractional lag.

2. Recognize that regression coefficients of lagged explanatory variables act as weights and that with adjacent lags, such as X_{t-2} and X_{t-3}, equal regression coefficients yield approximately the same result as a 2½-quarter lag. When only quarterly data are available, working with regression coefficients as weights is the only alternative. The economic interpretation of equal regression coefficients (weights) for X_{t-2} and X_{t-3} is that half of the purchasers buy 2 quarters after the explanatory variable occurs and half of the purchasers buy after 3 quarters. Unequal weights yield other interpretations, and these conceivably could be verified, or predetermined, by surveys of purchasers.

A second example of finding a single-integral time lag appears in Section 9.5 in the Process Control case study.

9.4 Distributed Lags: Geometric, "V," Almon, and Others

A more general functional form for Equation 9.2 is the following form which allows the current and many lagged values of X_t to affect Y_t:

$$Y_{ct} = b_1 + b_2X_{2t} + b_3X_{2(t-1)} + b_4X_{2(t-2)} + \ldots + b_mX_{2(t-m)} \qquad (9.7)$$

This regression equation illustrates a *distributed lag model* because the influence of the explanatory variable on Y_{ct} is distributed over a number of lagged values of X_t.[2]

Distributed lags consist of regression equations of the form of Equation 9.7, with the additional condition that the theoretical distribution of the b_i regression coefficients follow a predetermined form or shape. One common shape is that of the *geometric distribution*, where each b_i value declines by a given percentage at each additional lag. Another common form, the inverted V, requires that the b_i coefficients increase up to a point and decrease thereafter.

Distributed lag models, where the shape of the b_i coefficients is specified, are necessary for long lag distributions, such as for electric power generating plants that may require as many as forty quarterly lags or ten years. In a distributed lag model, only the parameters of the assumed lag distribution need be counted as degrees of freedom lost, whereas with individually fitted regression coefficients, each lag "costs" one degree of freedom. Five types of theoretical lag distributions are illustrated in Figure 9.4—the Jorgenson, Almon, Deleeuw (inverted V), Grilliches-Wallace, and Evans (double inverted B).

Calculating distributed lags requires computer programs with sophistication beyond least-squares linear regression routines and is beyond the scope of this discussion. See M.K. Evans, *Macroeconomic Activity: Theory, Forecasting, and Control*, which is particularly useful in describing applications of distributed lags (see Bibliography at the end of this chapter).

9.5 Dummy Variables and Time Trends

A dummy variable is an abstract number or set of numbers used as an explanatory variable in a multiple regression predicting equation. The "abstract" characteristic differentiates a dummy variable from a "real" variable, which is based on measurements of actual activity, such as construction contracts.

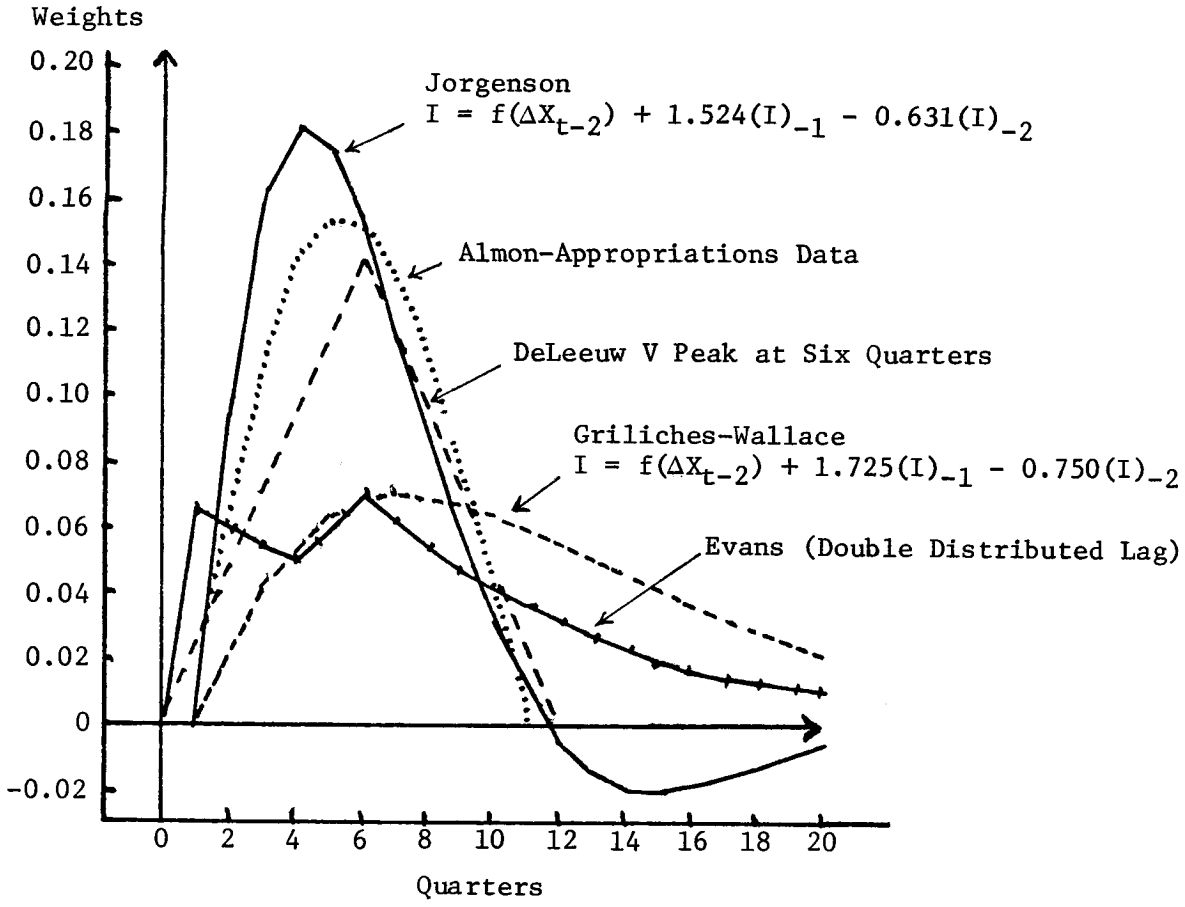
Table 9.3 lists the common types and forms of dummy variables. See Draper and Smith, *Applied Regression Analysis*, for an excellent discussion of methods of using dummy variables.

9.5.1 Single-Period Dummy Variable

The single-period dummy variable is used to represent an unusual economic circumstance, as when a company's sales are unusual due primarily to an employee strike, and when all of the effect is confined to one period. An example appears in Figure 9.5 showing quarterly seasonally adjusted sales of Delta Air Lines, Inc. Delta sales reflect largely domestic passenger and cargo operations. The sharp *increase* in the third quarter of 1966 was due to an employee strike on five other airlines which forced additional air traffic to Delta.

The most convenient form of dummy variable for a one-period effect uses "unity" in the period affected and "zero" for all other periods. The least-squares regression method will assign to the dummy variable all of the residual error after taking into account the normal influence of other explanatory variables in the equation. The "unit" form of dummy variable is convenient because its regression coefficient yields a handy quantitative measure of the influence of the unusual economic circumstance, here the strike.

Figure 9.4
Types of Theoretical Lag Distributions
Net Investment-Time Response to an Increase in Sales



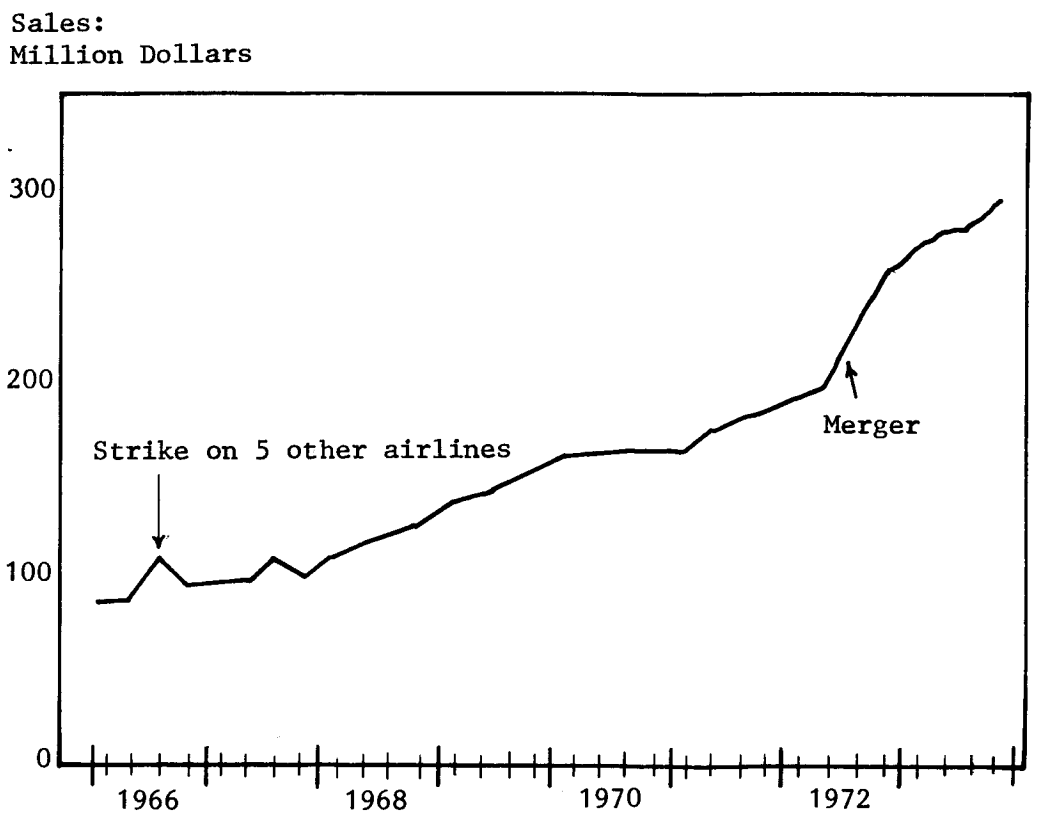
Source: Michael K. Evans, Macroeconomic Activity: Theory, Forecasting, and Control (New York, Harper & Row, 1969), p. 104.

Table 9.3 Dummy Explanatory Variables

Type	Example of Form
a. Single period	1 in period; 0 elsewhere
b. Multiple period "neutral"	+.5 in period preceeding unusual event such as a strike -1.0 in period of event +.5 in period succeeding event 0 elsewhere
c. Change in level	0,0,0,0,1,1,1,1
d. Time trend	1,2,3,4,5, . . .
e. Change in slope of linear time trend	0,0,0,0,1,2,3,4,5 or -5,-4,-3,-2,-1,0,0,0,0

Figure 9.5

Delta Air Lines, Inc., Sales: Seasonally Adjusted



Ideally, the forecaster should derive a direct measure of the influence of the unusual economic circumstance and use this measure as an explanatory variable. For a strike, this measure might be estimated as the number of employees involved times the number of days on strike times the net sales effect per day per employee. But such data are usually difficult or impossible to obtain because of the many unusual conditions present in a strike situation.

In practice, therefore, if an unusual economic circumstance can be clearly identified as a causal influence, then professional forecasters use a dummy variable. If the strike lasted only one quarter when forty quarters were being studied, the use of one degree of freedom is little detriment. In the Delta case, the regression coefficient of the dummy variable was positive and highly significant, with a t statistic of 4.37.

The *fundamental advantage* in using a dummy variable is that it allows other explanatory variables to receive regression coefficients that reflect their normal relationship to sales in all periods other than the period in which the dummy variable applies. Therefore, the other explanatory variables will "fit better." Using a dummy variable for a legitimate unusual circumstance is always preferable to omitting the unusual period because a gap in time-series regression may lead to error.

The single-period dummy variable assumes that all of the effect is in one time period and that this effect, either positive or negative, is not anticipated in advance or recovered in subsequent periods. This concentration of all effects in one quarter tends to be true in service industries but is less true in durable goods industries.

9.5.2 Multiple-Period Dummy Variable

Many unusual economic circumstances are not confined to one quarter and, therefore, a multiple-period dummy variable may be necessary. A special case is the multiple-period-neutral dummy variable which assumes, for example, that sales are not lost to competition during a strike and can be made up equally before the strike and after the strike. This was partly true in the U.S. steel industry strike 25 years ago. The first "+.5" part of the dummy variable in Line C of Table 9.3 represents an addition to sales caused by customers stockpiling steel before the date of an expected strike; the "−1.0" represents sales lost during the quarter of the strike; and the second "+.5" represents catch-up sales after the steel strike. The dummy variable numbers were arbitrary assumptions here. To the extent possible, dummy variable numbers should be based on measured estimates of effects, usually from previous similar occurrences.

Other types of multiple-period dummy variables than

the special "neutral" one discussed previously are possible. We give no illustrations here, and we caution that as dummy variables are lengthened in an unstructured manner, additional degrees of freedom are used up and economic causation may be more difficult to establish.

9.5.3 Change-in-Level Dummy Variable

The change-in-level dummy variable is appropriate when a major change is expected to represent a continuing condition, such as when Delta Air Lines, Inc. acquired the routes formerly operated by Northeast Airlines. In this case a dummy variable of Type C from Table 9.3 was added to the forecasting model, effective in the third quarter of 1972. Note the sharp increase in sales in Figure 9.5. The dummy variable allows a statistical measure of the net change in level of sales after the merger but allows for the normal effect of the other causal variables. The regression coefficient for this dummy variable was also positive and highly significant for Delta, with a t statistic of 9.55.

9.5.4 Time-Trend Dummy Variable

The time-trend type of dummy variable is a convenient way of determining a linear time trend in a multiple regression program. Any set of numbers with a constant change from one period to another would satisfy the requirement mathematically. But the form in Table 9.3 with a unit increase per period is convenient in that the regression coefficient times this time trend has an easily determined meaning, namely, the average increase per period in the dependent variable that is not attributable to other independent variables.

9.5.5 Change-In-Slope Dummy Variable

The change-in-slope-of-linear-relation form of dummy variable is handy when a new rate of change over time is obviously necessary because of a changed set of economic influences. This type of dummy variable is used in connection with a time trend to fit the two linear trends in the Safeway annual data in Section 3.4, Figure 3.1.

9.6 Transformations of Variables

A transformation of a variable in regression analysis is a change in the algebraic method of expressing the variable, such as transforming X_2 to $\log X_2$. Transformations may be made on the sales variable, or on one or more explanatory variables, or on both. The *purpose* of a transformation is to express the variable in a way that more accurately describes the underlying theoretical causal relationship, or that better fits the observed data relationships, or both.

Ideally, transformations should be made at the start of a regression study to accurately express known causal relationships or known empirical relationships among data. But many transformations are made after a regression study has started because preliminary regression results are not adequate, or do not conform to the assumptions necessary for significance testing, or both. We have called the analysis of observed regression relationships "diagnosis" in the sense of determining what is wrong at a given stage of analysis, regardless of why, and determining the most logical next steps to take to make and test improvements.

This section will describe the following four classes of transformations:

1. Algebraic transformations which change the original absolute form of a variable to a different algebraic form, such as transforming X_2 to $\log X_2$ or to X_2^2.
2. Economic transformations which alter the economic concept of a variable, such as transforming a current-dollar variable to a constant-dollar variable.
3. First-difference transformations, which express variables as changes from previous values rather than as absolute values.
4. Nonlinear transformations, which are extensions of the foregoing algebraic transformations. Nonlinear transformations are needed to express certain functions, but nonlinear transformations require different and more complex solution methods than the first three classes.

When considering transformations, the economic or business causation underlying the selection of explanatory variables has presumably already been established, meaning that the researcher has accepted that variable X_i has a causal relation to Y. The objective in studying transformations, then, is generally, not to find new economic concepts to be represented by the explanatory variables, but rather to find that form of an equation connecting sales and the one or more explanatory variables that will most closely represent the relationship. "Most closely" means that relationship having the smallest standard error of the sales variable or largest R^2, subject to the usual concerns of wanting nonautocorrelated residuals, homoscedasticity in the residuals, and so on. The search for a better form of the regression equation through transformations, however, may affect the final choice of explanatory variables to be included in a regression equation.

The effect of a transformation may be small or very extensive. For example, changing from the absolute form of the explanatory variable X_2 to the log of X_2 will only change the shape of the regression function from a straight line to a curve. However, shifting from the absolute form to a first difference equation causes a major change in the measurement concept and in the interpretation of the final equation, and it also may affect substantially the choice of explanatory variable.

To summarize this introduction to transformations, we repeat that the *purpose* of making transformations is to improve the regression relation for any of several reasons. The *type* of transformation to make will be illustrated in subsequent descriptions. Some of the indications of *need* for a transformation were given in Section 8.1 under the assumptions for regression and in Section 8.2 under the "Plot of Residuals" where testing for linearity and homoscedasticity were described. Further indications of the need for transformations will be given.

9.6.1 Algebraic Transformations

Algebraic transformations change the mathematical form of the regression predicting equation. Table 9.4 shows three types of algebraic transformations and examples of each.

Table 9.4 Algebraic Transformations

Type	Equation form		
Logarithmic	$Y_c = b_1 + b_2 \log X$	or	(9.8)
	$\log Y_c = b_1 + b_2 X$	or	(9.9)
	$\log Y_c = b_1 + b_2 \log X$		(9.10)
Polynomial	$Y_c = b_1 + b_2 X^2$	or	(9.11)
	$Y_c = b_1 + b_2 X^{\frac{1}{2}}$	or	(9.12)
	$Y_c = b_1 + b_2 X + b_3 X^2$		(9.13)
Reciprocal	$Y_c = b_1 + \frac{b_2}{X}$	or	(9.14)
	$\frac{1}{Y} = b + \frac{b_2}{X}$		(9.15)

We call these "algebraic transformations" because the need for them usually arises from mathematical reasons rather than from economic causation reasons. We recognize, however, that any transformation causes a mathematical change in the regression equation, regardless of the reason.

A *logarithmic transformation* of a variable consists of substituting the logarithm of a variable for the variable. This has the effect of substituting a measure of relative change of the variable for the absolute change. For example, in Table 9.4, the first type of logarithmic transformation, in Equation 9.8, shows the substitution of *log X* for *X*. Then any change in the sales prediction, Y_c, depends on a relative change in *X* rather than on an absolute change in *X*. Thus logarithmic transformations frequently fit population variables well.

The graphic effect of a logarithmic transformation is to condense the scale for large values of the transformed variable. Therefore, the logarithmic transformation is often helpful when the plot of residuals, or $(Y - Y_c)$, shows increasing dispersion with increasing values of the explanatory variable.

This can also be seen directly in a scatter diagram when the dispersion of *Y* values around the Y_c line increases for increasing size of the explanatory variable. The increasing size of the explanatory variable frequently reflects general economic growth over time, with the most recent observations at the upper right hand side of the scattergram. The increasing pattern of residuals may reflect a percentage variation rather than absolute variation, and the logarithmic transformation may significantly improve the R^2.

The logarithmic transformation provides one kind of nonlinear regression function, in the sense that the line of regression is curved for two-variable regression. The logarithmic transformation causes little difficulty in calculation, however, because most least-squares multiple regression computer programs accept the transformed numbers instead of the original numbers. The regression calculations will be valid so long as the transformations are within the class known as "intrinsically linear," which will be described as the last type in this section.

The *polynomial* transformations illustrated in Table 9.4, Equations 9.10, 9.11, and 9.12, are most likely to be useful where the pattern of dispersion in the residuals indicates a *systematic* deviation from uniform scatter. For example, if the scale of the X variable needs to be expanded for large values of X in order to secure a more uniform dispersion of residuals, then transforming X to X^2 may help, as in Equation 9.11. If the X variable needs to be contracted, try transforming X to $X^{\frac{1}{2}}$, as in Equation 9.12. If the residuals tend to be high at the left and right sides of the residual chart and low at the center, or vice versa, then polynomial transformation like Equation 9.13 is likely to be helpful.

The *reciprocal* transformation consists of substituting 1/X for X, as an example. The result is that when X increases (after being transformed to 1/X), then the predicted value of Y decreases, but by decreasing relative amounts as X increases. The effect is, therefore, not the same as changing the sign of the regression coefficient. The reciprocal transformation of X is particularly helpful where the direction of influence on Y_c is to be reversed and where the scale of X is to be contracted for large values of X, all in the special nonlinear way characteristic of reciprocals.[3]

If several algebraic transformations yield homogeneity,

then use considerations of economic theory to choose the preferred transformation. If several transformations are similar in effect, then the logarithmic transformation (reflecting approximately constant variations in the percentage residuals) is the most straightforward transformation to explain to other people because of its similarity to percentage growth rates and compound interest.

9.6.2 Economic Transformations

We have called the following transformations "economic" transformations because the stimulus for their use comes primarily from economic considerations. Many of the transformations use simple algebra. Table 9.5 lists the types and the equation forms or examples.

The *constant dollar* transformation converts a current dollar time series to a constant dollar time series. In general, a dependent sales variable in physical units or constant dollars must always be correlated with explanatory variables in physical units and/or constant dollars. The dependent sales variable in current dollars may, in simple cases, be correctly regressed against explanatory variables in current dollars, and this has been done in several cases in this book. Use of current dollar sales and explanatory variables may be justified when dollar sales of products or services are considered a function of current dollar income or when both the sales and the explanatory variables are subject to approximately the same rates of price change or inflation.

Many regression equations show a constant dollar dependent variable as a function of a constant dollar explanatory variable and of a price ratio. The objective here is to have the price ratio be independent of the other variables, and this requires avoiding double counting of price in the ratio and in the other variable. Using constant dollars avoids the double counting of price. Econometric work largely uses constant dollar and physical unit variables, along with price deflator series, price ratios, and price differences. The price deflators are used to convert

Table 9.5
Economic Transformations

Type	Equation form or example
Constant dollar	$\frac{\text{Current dollar time series}}{\text{Price deflator}}$
Ratio to another variable	$\frac{\text{Product price}}{\text{Competitive product price}}$
	$\frac{\text{Aggregate income}}{\text{Population}} = \text{Per capita income}$
	$\frac{\text{Aggregate income/price index}}{\text{Population}} = \text{Real per capita income}$
Ratio to trend	$\frac{\text{Time series}}{\text{Trend of time series}}$
Change in level of aggregation	Change from Total Construction Contracts to Residential Construction Contracts (example of disaggregation)
	Change from Personal Disposable Income to Personal Income (example of aggregation, accomplished by adding elements of personal taxes and personal savings.)

current dollars to constant dollars for comparison with other constant data series. The price deflators themselves are often determined by simultaneous equation solutions within the econometric model.

The constant dollar transformations were illustrated in Section 9.2 for industrial wheel tractors and construction contracts, where current dollar construction contracts were divided by a construction price deflator to arrive at a constant dollar value of construction contracts, for use with industrial wheel tractor sales in physical units.

To summarize, the main reason for transforming current dollar data to constant dollars is to provide comparability where different rates of price inflation apply to the original variables. This frequently happens. A related reason is that many economic theories are expressed in constant dollar or "real" terms to reflect purchasing power influences.

The *ratio-to-another variable* transformation, as shown in Table 9.5, is necessary wherever the ratio of two variables has a special meaning, separate from or in addition to their meaning in absolute form. The ratio of two prices are frequently used as an explanatory variable in demand functions. Another common use of ratios is to convert aggregate income into per capita income.

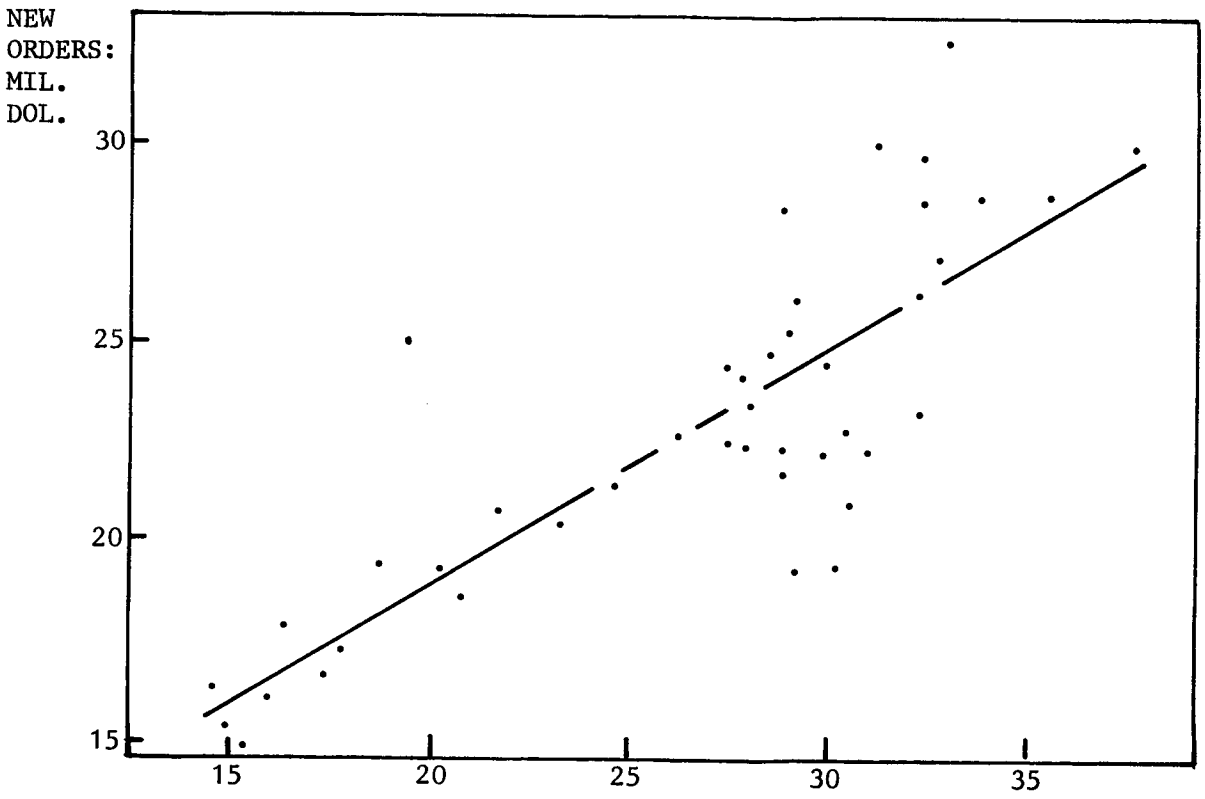
A further use of the ratio-to-another-variable transformation will be illustrated in Chapter 14 on preparing company and industry forecasts. There we use a ratio of company sales to industry sales. The resulting ratio, called "C%I," is a market share variable with a very special meaning.

The *ratio-to-trend* transformation is highly useful where the sales and explanatory variables both have substantial business cycle fluctuations but where the amplitude of cyclical fluctuations may be different. In this case, expressing both variables as ratios to their own time trend provides a way of calculating the relative amplitude of fluctuations around trend. Then the least-squares regression calculations can adjust for these varying amplitudes.

Figure 9.6 shows the scattergram for the data of Figure 9.3. Note that a linear regression line would fit reasonably well at the lower left of the bivariate observations and at the upper right for the two observations near the right margin. But within the observations centering around the vertical ordinate for NPEE (mfr) of $30 billion, the observations had a wide scatter. The plot with points connected in time shown in Figure 7.4 shows that the movements through time around the $30 billion ordinate are nearly vertical, rather than near the sloping least-squares regression line. This indicates that the present regression model is *not* fitting well during the sharp up and down movements of these two cyclical series. From an analytic standpoint, we have the wrong explanatory variable, the wrong functional form, or parts of both.

Figure 9.6

Scatter Diagram: Process Control and NPEE (T+2)



A review of economic causation confirms that business expenditures for new plant and equipment clearly encompass the market for Process Control products, that a measure as broad as NPEE (Mfr) is the correct independent variable, and that it is at about the correct level of aggregation.

Figure 9.7 shows the actual and predicted values from linear simple regressions for the sample (or historical) period. Note that the cyclical peaks of the predicted line in 1966 and 1969 almost correspond to the cyclical peaks of the actual and that the cyclical trough of the predicted line in 1970 corresponds to the trough in the actual line. Both series rise together in 1972. This suggests that concurrent timing of peaks and troughs is not the problem. The problem obviously is that predicted sales have less vertical amplitude of variation than the actual sales.

A related critical problem is that no business cycles appear during 1962 to 1965. Hence, any function that fits well into 1962 through 1975 may not fit well in 1966 through 1972, and conversely any function that fits well cyclically in 1966-1972 may not fit well in 1962-1965. Thus a choice has to be made, and clearly the old data for 1962-1965 are less relevant for forecasting than the recent data for 1966-1972. Therefore, the first three years are dropped to allow the use of a function that magnifies the amplitude of the business cycle fluctuation of NPEE $(Mfr)_{t+2}$ in 1966-1972.

A ratio-to-trend transformation, together with omission of the 1962-1965 data, allows for a magnification of the NPEE cyclical amplitude of variation. The resulting predicting equation is:

$$\frac{\text{Predicted sales}}{\text{Sales trend}} = b_1 + b_2 \frac{\text{NPEE (Mfr)}_{t+2}}{\text{NPEE TREND}} \qquad (9.16)$$

or

$$Y_c = 0.9374 + 1.9372X_{2(t+2)} \qquad (9.17)$$

Figure 9.8 shows the predicted values graphically. Note that the amplitude of the business cycle's variations of NPEE is now much greater and roughly matches that of the actual. The $R^2 = .81$, which is appreciably higher than $R^2 = .54$ of the absolute values in the 1962-1972 data. This is a much more useful equation for forecasting because the predicted values fit much better since 1966. Tests of forecast Equation 9.17 and comparisons of actual with forecast are given in Chapter 15.

The last type of economic transformation listed in Table 9.5, the *change-in-level-of-aggregation* transformation, either subdivides a time-series variable into two or more economic components or adds two or more time-series variables to derive a new variable. This transformation

Figure 9.7

Process Control and $NPEE_{T+2}$: Actual Versus Predicted Sales for 1962-72

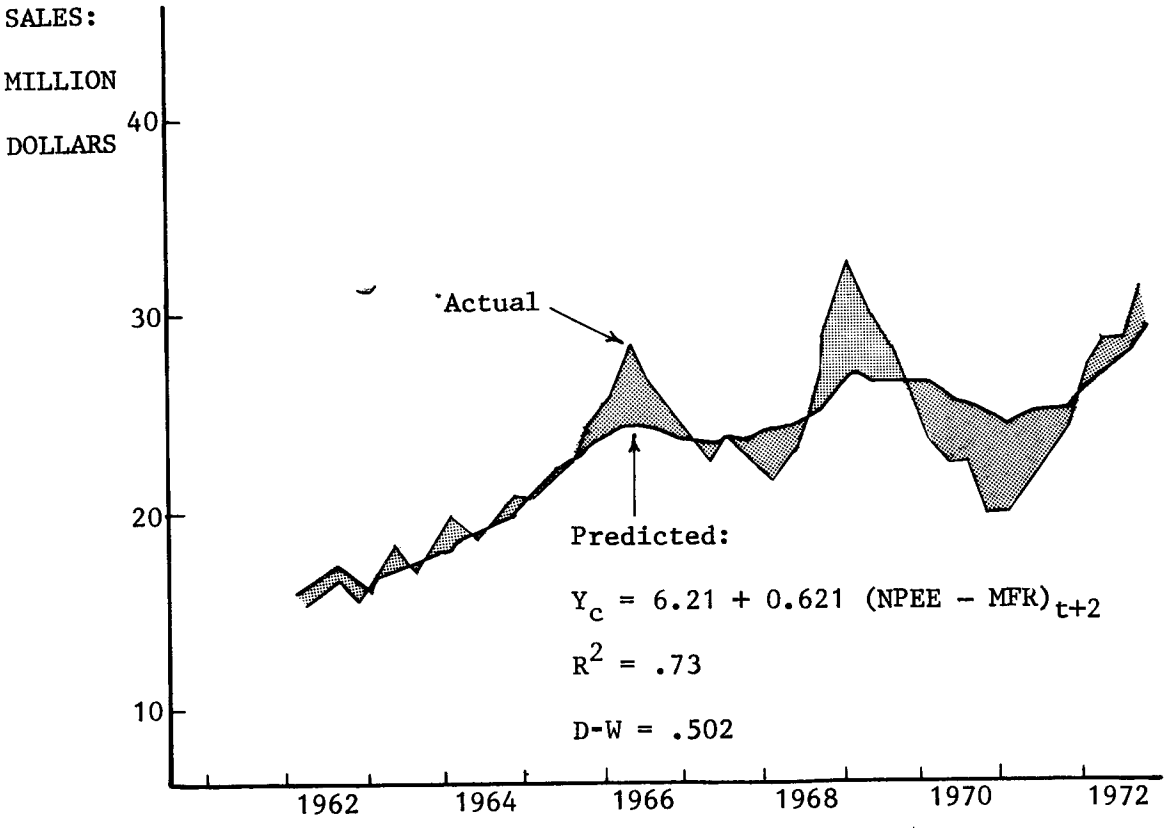
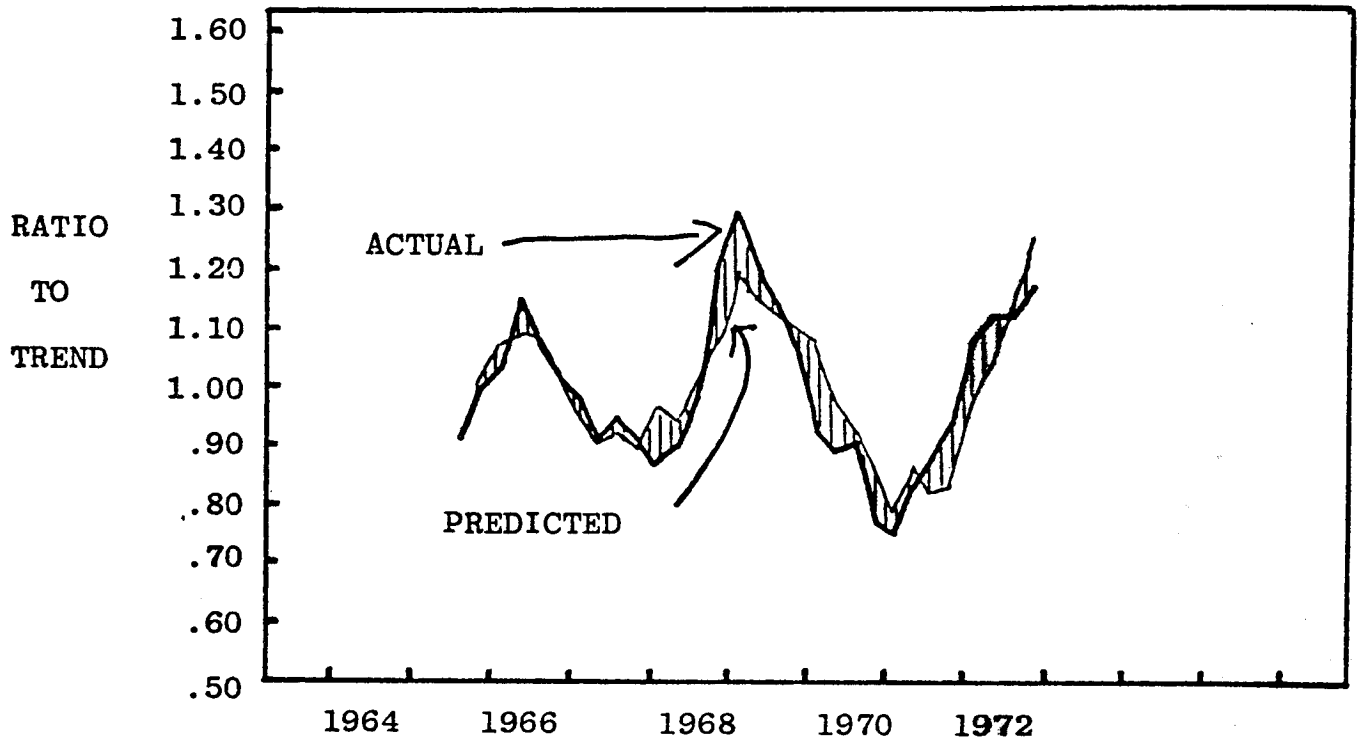


Figure 9.8

PROCESS CONTROL AND $NPEE_{T+2}$ WITH RATIO-TO-TREND TRANSFORMATIONS

ACTUAL VS PREDICTED FOR 1965-72



involves a change in numbers, but the change yields both a strictly mathematical effect and a change-in-economic-concept effect. The change-in-level-of-aggregation transformation will be illustrated next in a test of whether the appropriate *level* of aggregation has been used when the correct general economic concept is known.

To illustrate with the Process Control Company case, suppose the forecaster first started to correlate Process Control Sales with Total New Plant and Equipment Expenditures. To see the alternatives, all categories of NPEE published in the *Survey of Current Business* appear in Table 9.6. The forecaster for Process Control might determine from the marketing department that, although the majority of Process Control products go to producers of machinery for manufacturing industries, a small percentage of products also go to producers of machinery for nonmanufacturing industries, e.g., utilities. Thus a forecaster might have started by regressing Process Control sales with Total New Plant and Equipment Expenditures at the level of $88.44 billion in 1972.

If the statistical results are poor, then the forecaster might restudy the market to which Process Control goods are shipped and may concentrate on manufacturing industries, which represent about 90 percent of the end uses of Process Control products.

This process of moving from a time series measuring total activity in a particular sector of the economy to one or more subdivisions of that particular sector is called disaggregation. Early work with Process Control went through this process, which led to the conclusion that "manufacturing" was the best level of disaggregation. A further possible disaggregation, if justified by new data on end uses of Process Control products, might be to subdivide the manufacturing NPEE series into durable goods industries and nondurable goods industries.

When we disaggregate by subdividing a time series into two or more components, we are nevertheless shifting to a different economic concept. Disaggregating (or its reverse, aggregating), therefore, is not purely algebraic. An explanatory variable at a different level of disaggregation may have business cycle peaks at different times, may have different amplitudes of fluctuation over the cycle, and may have other differences.

How can we tell if disaggregation will help? Suppose we consider moving down one level of aggregation from manufacturing to two subdivisions: durable goods manufacturing and nondurable goods manufacturing. If durable goods manufacturing constituted a precisely uniform percent of total manufacturing, and thus represented "total manufacturing" multipled by a constant less than unity, then the result of this disaggregation would *not* help build an improved regression equation. On the

Table 9.6

New Plant and Equipment Expenditures, United States, 1972

Industry	1972 Expenditures	
	Billion Dollars	
Durable goods manufacturing	15.64	
Nondurable goods industries	15.72	
Manufacturing total		31.35
Mining	2.42	
Railroad	1.80	
Air transportation	2.46	
Other transportation	1.46	
Public utilities	17.00	
Communication	11.89	
Commercial and other	20.07	
Nonmanufacturing total		57.09
Total		88.44

Source: Survey of Current Business, U.S. Department of Commerce, Vol. 54, No. 2, Feb. 1974, p. S-2.

other hand, if durable goods manufacturing represents a *varying* percentage of total manufacturing, then disaggregating to durable goods manufacturing represents moving to a different concept and may improve the regression.

Conversely, the process of aggregation, such as moving from just manufacturing NPEE to total NPEE, is also not a simple algebraic transformation because, in general, manufacturing NPEE would not represent a uniform proportion of total NPEE.

The choice among different levels of aggregation must be based primarily on economic causation, that is, on matching the sector of economic activity of the explanatory variable as closely as possible to the end uses of the products or services in the sales variable.

Disaggregation applies as well to the dependent sales variable as to an explanatory variable. For example, if Process Control products could be subdivided by end uses in manufacturing industries versus nonmanufacturing industries, then conceivably the forecast operation could be divided into one regression equation for manufacturing and another for nonmanufacturing. Then the total Process Control sales forecast is the sum of the two sales categories. If such subdivision can be made, improved regression equations usually may be found.

Other examples of disaggregating an explanatory variable include:

1. Disaggregating personal disposable income into personal consumption expenditures and personal savings. These series have important differences in economic concept.

2. Disaggregating the labor force into employed persons plus unemployed persons. Again, these two subdivisions behave quite differently, with the unemployed sector showing far greater sensitivity to business cycle changes.

Disaggregating an explanatory variable economic time series into subcomponents as illustrated previously frequently improves the regression (increases the R^2, and so on) for these reasons:

1. Two explanatory variables provide a further degree of *statistical flexibility* in fitting a regression equation than one explanatory variable. A higher R^2 will usually occur but not if good economic reasoning has been violated.

2. The disaggregated explanatory variable series may match the general *economic concept* of the sales variable more closely than the aggregate. If a forecaster can determine the correct level of aggregation by economically based reasoning at the start of the project, he should do so. But sometimes this determination is not made in order to get a "quick look at the data," or in less obvious cases may require substantial information on end uses of sales which may not be available at the start of a project.

3. The *company characteristics* of the products or services and of the marketing organization may result in a much better regression equation when using two or more disaggregated explanatory variables with company sales than when using a single aggregate variable. This is an extension of the more precise "economic concept" from paragraph 2, but it emphasizes company characteristics rather than general economic class characteristics of the sales variable.

9.6.3 First Difference Transformations

The first difference transformation is highly useful when the explanatory variable is a slowly changing aggregate, such as U.S. personal income. The first difference transformation amplifies the influence of changes in the aggregate from one period to another by making the entire equation depend on changes, called first differences, rather than on the absolute value of the aggregate. The algebraic symbol for the first difference is Δ, the Greek letter "delta." But expressing a variable as a first difference fundamentally changes the nature of the regression equation and also fundamentally changes the meaning of the diagnostic measures, such as the standard error of regression, the R^2, and the Durbin-Watson test for autocorrelation.

Let us begin by considering the standard linear regression equation for absolute values of Y and X:

$$Y_{ct} = b_1 + b_2 X_t \tag{9.18}$$

Here we will consider t as the first forecast period. When we transform X to ΔX, then Equation 9.18 must be transformed to

$$Y_{ct} = b_1 + b_2 (\Delta X_t) + b_3 Y_{t-1} \tag{9.19}$$

Equation 9.19 can also be expressed in the residual form as follows, and we will use the latter for explaining the first difference transformation:

$$Y_t = b_1 + b_2(\Delta X_t) + b_3 Y_{t-1} + \epsilon_t \tag{9.20}$$

This equation has a "lagged dependent variable," Y_{t-1}, in the position of an explanatory variable because it is on the right-hand side of the equation. From the standpoint of time, "t," the value of "Y_{t-1}" is a predetermined or already known variable and, therefore, may be legitimately included on the right-hand side with the other independent explanatory variable. Notice now that the difference between Y_t on the left-hand and Y_{t-1} on the right-hand side in Equation 9.20 is accounted for by three additive terms:

1. The constant term, b_1, reflecting the average additive growth increment in Y_t per period *exclusive* of the effect of all other additive terms in the equation.

2. The product of the regression coefficient, b_2, times the first difference of the explanatory variable, ΔX_t.

3. The product of the b_3 regression coefficient times the lagged dependent sales variable. The b_3 should be thought of as a term from compound interest formulas, that is, $b_3 = (1+r)$, where r is the multiplicative growth rate per period in Y, exclusive of the effect of b_1 and $b_2\Delta X_t$. The ϵ_t term will have no average effect on Y_t or Y_{ct} because $E(\epsilon_t) = 0$. The existence of Equation 9.20 implies another underlying equation:

$$\Delta Y_t = b_1 + b_2 \Delta X_t + \epsilon_t \tag{9.21}$$

You might start by searching for explanatory variables that

are expressed in first difference form and then calculating the correlations among the ΔY and ΔX_i values. You would usually choose that ΔX_i series yielding the highest correlation. The important point here is that searching in this way would lead to the same result as searching for explanatory variables in Equation 9.19, and searching through trials of Equation 9.19 is usually more convenient for purposes of analyzing results.

Equations of the form of 9.19 and 9.20 are called *recursive* equations, meaning that these equations forecast only one period at a time for successive periods. To prepare a forecast for $Y_{c(t+4)}$, for example, requires first preparing a forecast Y_{ct} using the predetermined lagged dependent variable, Y_{t-1}, plus b_1 and $b_2\Delta X_t$. Then forecast $Y_{c(t+1)}$ using $b_1 + b_2\Delta X_{(t+1)} + b_3 Y_{ct}$, then successively and finally, $Y_{c(t+4)}$. By contrast, the forecast for $Y_{c(t+4)}$ in Equation 9.18 could be calculated directly, given a forecast of the explanatory variable X_{t+4}. This characteristic of recursive equations is normally not a problem in applying multiple regression analysis because computer programs to handle recursive equations are widely available and because forecasts of explanatory variables for all periods in a forecast span are normally prepared and serve as data inputs to generate forecasts.

The major difficulty in recursive equations arises because the standard error of regression and the R^2 must be interpreted differently than for an equation with all absolute values of variables, like Equation 9.18. The standard error of regression in the usual equation with absolute variables measures the average error of predicted sales from actual sales for *all* historical periods, but, by contrast, the standard error of regression for a recursive equation measures only the average error from a previous actual value, Y_{t-1}, to the next predicted value, Y_{ct}. Thus the recursive standard error is an average error for all spans of *one* period each.

If the recursive standard error is used for a forecast five periods ahead, then whatever confidence coefficient is calculated for a one-period forecast confidence interval would have to be expressed as a power of the number of periods for a multi-period forecast. Thus, if a 0.95 confidence coefficient applies to a specific confidence interval, then for a five-period recursive forecast the applicable confidence coefficient is approximately $(0.95)^5 = (0.77)$. We say "approximately" because other elements of forecast error also apply, as outlined in Chapter 15.

Similarly, the R^2 in a recursive equation measures the ratio of explained to total variance where the total variance reflects errors in historical predictions only from period to period. By contrast, the R^2 for absolute data as in Equation 9.18 measures the ratio of explained to total variance where total variance measures the average errors in all historical predictions.

The meaning and significance testing of the Durbin-Watson test for autocorrelation of the residuals also changes in a recursive equation. The Durbin-Watson test was designed to measure autocorrelation in residuals from absolute regression equations like Equation 9.18. If the Durbin-Watson "d" is calculated for recursive predictions, it will still measure autocorrelation of residuals, but narrower confidence limits must be used for the significance test. The reason is that since recursive equations measure changes from one period to a successive period only, then measures of dispersion are smaller.

9.6.4 Nonlinear Transformations

The need occasionally arises for algebraic transformations that expand or contract the variations of a variable in particular ranges of that variable or for complex functional forms of the regression equation indicated by definite information about the form of the relationship. These needs may call for algebraic transformations beyond those discussed earlier and particularly for a class of transformations called *intrinsically nonlinear* transformations.

The transformations listed in Section 9.6.1 are all *linear in the parameters* and are of the type

$$Y = B_1 + B_2Z_2 + B_3Z_3 + \ldots + B_pX_p + \epsilon \qquad (9.21)$$

where the Z_i can represent any function such as X^2, $1/X$, and so on, of the basic explanatory variables $X_2, X_3, \ldots, X_k$.

Equation 9.21 is *linear in the parameters*. Draper and Smith call the form of Equation 9.21 *intrinsically linear.*[4] This intrinsically linear characteristic is important because, if present, most multiple regression computer programs will readily carry out the complex regression computation.

The estimation of the parameters in an *intrinsically nonlinear* regression function is not a straightforward calculation as it is with intrinsically linear functions. Solving intrinsically nonlinear regression problems usually requires an iterative technique and special computer programs.

Footnotes

1. Robert A. Berman, "An Econometrician's Reaction to a Non-econometrician's Guide to Econometrics," *Business Economics*, Vol. IX, No. 1, January 1974, p. 81, where he states, "Multicollinearity may understate the errors of the regression coefficients, but its presence will not 'produce spurious forecasts.' Multicollinearity becomes a problem only in identifying particular coefficients in an equation for the purpose of testing certain hypotheses about them. It is *not* a problem when one wishes simply to forecast."

2. Jan Kmenta, *Elements of Econometrics* (New York, The Macmillan Company, 1971), p. 473.

3. Draper, Klingman, and Weber, *Mathematical Analysis: Business and Economic Applications* (New York, Harper and Row, 1972), provides a helpful section on p. 94, entitled "Applications of Nonlinear Curves in Business and Economics."

4. Draper and Smith, *Applied Regression Analysis* (New York, John Wiley and Sons, Inc., 1966), pp. 263-264 and 267.

Bibliography

Dhrymes, Phoebus J. *Distributed Lags: Problems of Estimation and Formulation*. San Francisco: Holden-Day, Inc., 1971. An excellent advanced text, ch. 9.

Draper, N.R. and H. Smith. *Applied Regression Analysis*. New York: John Wiley & Sons, Inc., 1966, ch. 5, p. 134.

Evans, Michael K. *Macroeconomic Activity: Theory, Forecasting and Control*. New York: Harper & Row, 1969, p. 95 and p. 204.

Huang, David S. *Regression and Econometric Methods*. New York: John Wiley & Sons, Inc., 1970, pp. 102, 135, 149-158, 163, 180.

Johnston, J. *Econometric Methods*. 2nd ed. New York: McGraw-Hill Book Company, 1972, ch. 10.

Kane, Edward J. *Economic Statistics and Econometrics*. New York: Harper & Row, 1968, pp. 277, 353, 364.

Kmenta, Jan. *Elements of Econometrics*. New York: Macmillan Company, 1971, pp. 380, 473.

Salzman, Lawrence. *Computerized Economic Analysis*. New York: McGraw-Hill Book Company, 1968, ch. 6.